



Sievertsen, H., Piovesan, M., & Gino, F. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences of the United States of America*, 113(10), 2621-2624. <https://doi.org/10.1073/pnas.1516947113>

Publisher's PDF, also known as Version of record

License (if available):  
Other

Link to published version (if available):  
[10.1073/pnas.1516947113](https://doi.org/10.1073/pnas.1516947113)

[Link to publication record in Explore Bristol Research](#)  
PDF-document

This is the final published version of the article (version of record). It first appeared online via PNAS at <http://www.pnas.org/content/113/10/2621>. Please refer to any applicable terms of use of the publisher.

## University of Bristol - Explore Bristol Research

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:  
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

# Cognitive fatigue influences students' performance on standardized tests

Hans Henrik Sievertsen<sup>a</sup>, Francesca Gino<sup>b,1</sup>, and Marco Piovesan<sup>c</sup>

<sup>a</sup>The Danish National Centre for Social Research, 1052 Copenhagen, Denmark; <sup>b</sup>Harvard Business School, Harvard University, Boston, MA 02163; and <sup>c</sup>Department of Economics, University of Copenhagen, 1353 Copenhagen, Denmark

Edited by Pamela Davis-Kean, University of Michigan, Ann Arbor, MI, and accepted by the Editorial Board January 15, 2016 (received for review August 25, 2015)

**Using test data for all children attending Danish public schools between school years 2009/10 and 2012/13, we examine how the time of the test affects performance. Test time is determined by the weekly class schedule and computer availability at the school. We find that, for every hour later in the day, test performance decreases by 0.9% of an SD (95% CI, 0.7–1.0%). However, a 20- to 30-minute break improves average test performance by 1.7% of an SD (95% CI, 1.2–2.2%). These findings have two important policy implications: First, cognitive fatigue should be taken into consideration when deciding on the length of the school day and the frequency and duration of breaks throughout the day. Second, school accountability systems should control for the influence of external factors on test scores.**

cognitive fatigue | time of day | breaks | standardized tests | education

**E**ducation plays an important role in societies across the globe. The knowledge and skills children acquire as they progress through school often constitute the basis of their success later in life. To evaluate the effectiveness of schooling on children and to provide data to better manage school systems and develop education curriculum, legislators and administrators across societies have used standardized tests as their primary tool as they commonly believe test data are a reliable indicator of student ability (1, 2). In fact, these tests have become an integral part of the education process and are often used in drafting education policy, such as the *No Child Left Behind Act* and *Race to the Top* in the United States. As a result, students, teachers, principals, and superintendents are increasingly being evaluated (and compensated) based on test results (2).

A typical standardized test assesses a student's knowledge base in an academic domain, such as science, reading, or mathematics. When taking a standardized test, the substance of the test, its administration, and scoring procedures are the same for all takers (3). Identical tests, with identical degrees of difficulty and identical grading methods, are propagated as the most fair, objective, and unbiased means of assessing how a student is progressing in her learning.

The widespread use of standardized testing is based on two fundamental assumptions (3): that standardized tests are designed objectively, without bias, and that they accurately assess a student's academic knowledge. Despite these goals in the creation of standardized tests, in this paper we identify one potential source of bias that drives test results and that is predictable based on psychological theory: the time at which students take the test. We use data from a context in which the timing of the test depends on the weekly class schedule and computer availability at the school and thus is random to the individual. These factors are common conditions of standardized testing. We suggest, and find, that the time at which students take tests affects their performance. Specifically, we argue that time of day influences students' test performance because, over the course of a regular day, students' mental resources get taxed. Thus, as the day wears on, students become increasingly fatigued and consequently more likely to underperform on a standardized

test. We also suggest, and find, that breaks allow students to recharge their mental resources, with benefits for their test scores.

We base these predictions on psychological research on cognitive fatigue, an increasingly common human condition that results from sustained cognitive engagement that taxes people's mental resources (4). Persistent cognitive fatigue has been shown to lead to burnout at work, lower motivation, increased distractibility, and poor information processing (5–12). In addition, cognitive fatigue is detrimental to individuals' judgments and decisions, even those of experts. For instance, in the context of repeated judicial judgments, judges are more likely to deny a prisoner's request and accept the status quo outcome as they advance through the sequence of cases without breaks on a given day (13). Evidence for the same type of decision fatigue has been found in other contexts, including consumers making choices among various alternatives (14) and physicians prescribing unnecessary antibiotics (15). Across these contexts, the overall demand of multiple decisions people face throughout the day on their cognitive resources erodes their ability to resist making easier and potentially inappropriate or bad decisions.

At the same time, research has highlighted the beneficial effects of breaks. Breaks help people recover physiologically from fatigue and thus serve a rejuvenating function (16, 17). For instance, workers who stretch physically during short breaks from data entry tasks have been found to perform better than those who do not take breaks (16). Breaks can also create the slack time necessary to identify new ideas or simply reflect (18–20), with benefit for performance.

In this paper, we build on this work by examining how cognitive fatigue influences students' performance on standardized tests. We use data on the full population of children in Danish public schools from school years between 2009/10 and 2012/13 (i.e.,

## Significance

**We identify one potential source of bias that influences children's performance on standardized tests and that is predictable based on psychological theory: the time at which students take the test. Using test data for all children attending Danish public schools between school years 2009/10 and 2012/13, we find that, for every hour later in the day, test scores decrease by 0.9% of an SD. In addition, a 20- to 30-minute break improves average test scores. Time of day affects students' test performance because, over the course of a regular day, students' mental resources get taxed. Thus, as the day wears on, students become increasingly fatigued and consequently more likely to underperform on a standardized test.**

Author contributions: H.H.S., F.G., and M.P. designed research; H.H.S., F.G., and M.P. performed research; H.H.S. analyzed data; and H.H.S., F.G., and M.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. P.D.-K. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

<sup>1</sup>To whom correspondence should be addressed. Email: fgino@hbs.edu.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516947113/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516947113/-DCSupplemental).

children aged 8–15) and focus on the effects of both time of the test and breaks—factors that directly relate to students' cognitive fatigue.

### The Study

In Denmark, compulsory schooling begins in August of the calendar year the child turns 6 and ends after 10 years of schooling. Approximately 80% of children attend public school (14% attend private schools and 6% attend boarding schools and other types of schools). With the purpose of contributing to the continuous evaluation and improvement of the public school system, in 2010, the Danish Government introduced a yearly national testing program called The National Tests. This program consists of 10 mandatory tests: a reading test every second year (grades 2, 4, 6, and 8), a math test in grades 3 and 6, and other tests on different topics (geography, physics, chemistry, and biology) in grades 7 and 8. Each test consists of three parts, presented in random order. (Importantly, there is no ordering of the subtests. The subareas are not tested after each other; rather, a student might first get a question to subarea 1, then to subarea 2, then to subarea 1 again, then to subarea 3, and so on.) For instance, the math test is divided into Numbers and Algebra, Geometry, and Applied Math. In our analyses, we take the simple average across these three parts and standardize the score by subject, test year, and grade (with mean 0 and SD 1). This approach enables us to interpret effects in terms of SD.

These tests are adaptive: the test system chooses the questions based on the student's level of proficiency as displayed during the test and calculates the test results automatically.

Our dataset comprises all two million tests taken in Denmark between school years 2009/2010 and 2012/2013. Data are provided by the Ministry for Education and linked to administrative registers from Statistics Denmark, a government agency. The administrative data give us information about sex, age, parental background (education and income), and birth weight. The parental characteristics are measured in the calendar year prior to the test year. Our sample consists of 2,034,964 observations from 2,105 schools and 570,376 students. We excluded 17,863 observations (0.9% of the initial sample) to ensure that only normal tests (i.e., tests that were not taken under special circumstances) were included (see *SI Text* for details). We made no other sample selection.

Two characteristics of these tests should be noted. First, the main purpose of these tests is for teachers covering specific topics (e.g., geography) to gain insight into each student's achievements for the creation of individually targeted teaching plans. Teachers have no obvious incentive to manipulate students' performance, and parents are presented with the test results on a simple five-point scale.

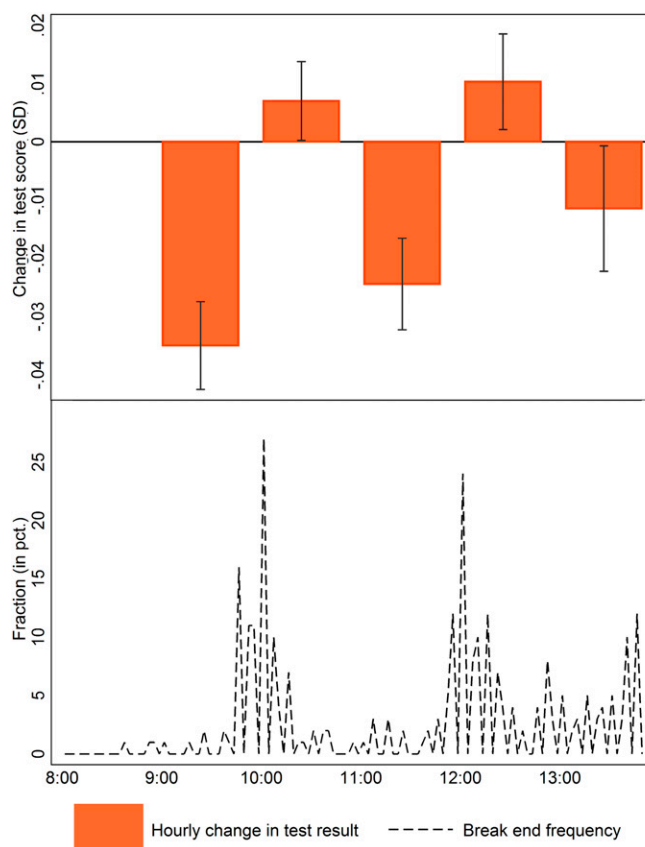
Second, these tests are computer based: to test the students, the teacher covering a specific topic has to prebook a test session within the test period (January–April of each year). Therefore, the test time is an exogenous variable because it depends on the availability of a computer room and students' class schedules. Our analysis confirms that students are allocated to different times randomly. In fact, covariates are balanced across test time, and our results are robust to using within-student variation (i.e., variation in test time across years within the same subject for the same student, as shown in *SI Text*). In short, our data represent a natural experiment and thus a unique opportunity to test the effects of time of day and breaks on test scores.

During the school day, students have two larger breaks during which they can eat, play, and chat. Usually these breaks are scheduled around 10:00 AM and 12:00 PM and last about 20–30 minutes. As we use a large sample of 2,105 schools, and each school can organize its schedule independently, we contacted 10% of the schools by phone and asked them about their breaks schedule. We received responses from 95 schools (a 45% response rate). Our interviews

revealed that 83% of the schools' first break starts between 9:20 AM and 10:00 AM and that 68% have a second break starting between 11:20 AM and 12:00 PM. Finally, we asked if test days follow a different schedule. Eighty-four percent of the schools we interviewed confirmed they follow the usual break schedule on test days. (Results using only the schools that we contacted confirm those reported below and are shown in *SI Text*.)

To test our main predictions, we first focus on the effect of test time. The upper panel of Fig. 1 shows the hour-to-hour difference in the average test score by test time. We created this graph by estimating a linear model of test score on indicators for test hour using ordinary least squares (OLS). In the model, we control for school, grade, subject, day of the week, and test-year fixed effects, as well as for parental education, parental income, birth weight, sex, spring child, and origin. As the graph shows, time of day influences test performance in a nonlinear way: although the average test score deteriorates from 8:00 to 9:00 AM, it improves from 9:00 to 10:00 AM. This alternating pattern of improvements and deterioration continues over the day (see *SI Text* for details).

Next, we focus on the effect of having the test after a typical break. By typical break, we mean a break that commonly occurs at the same time throughout the week, across schools. The dashed line in the lower part of Fig. 1 shows the breaks time. Breaks typically end just before 10:00 AM and 12:00 PM. Together, the



**Fig. 1.** Hour-to-hour effect on test scores and break patterns. Effects are estimated based on administrative data from Statistics Denmark. (*Upper*) How the average test score changes from hour to hour. (*Lower*) Distribution of when breaks end, based on a survey conducted on 10% of the schools. The hourly effect is estimated in a linear model controlling for unobserved time invariant fixed effects on grade, day of the week, and school level. We also control for test year fixed effects, as well as parental income, parental education, nonwestern origin, sex, spring child, and birth weight. The details on the model and estimation procedure are shown in *SI Text*, along with a table with regression results.

hour-to-hour changes and the break pattern show that test performance declines during the day but improves at test hours just after a break. Breaks, it appears, recharge students' cognitive energy, thus leading to better test scores.

Next, to provide further support for our hypotheses, we explicitly model the effects of time of day and breaks by estimating the linear relationship between test score, test hour, and breaks. The model is estimated by OLS and also includes the individual characteristics and the fixed effects described above. The point estimates on break and test hour are shown in Fig. 2. Fig. 2A shows these point estimates for various specifications and subsamples. The first two bars show that for the full sample, the test score is reduced by 0.9% (95% CI, 0.7–1.0%) of an SD for every hour (the red bar), but a break improves the test score by 1.7% (95% CI, 1.2–2.2%) of an SD (the blue bar). We then conduct the same analyses for various subsamples but find limited evidence of heterogeneous effect of breaks across subject (i.e., mathematics vs. reading) and age (i.e., young vs. old). For hour of the day, the effect is more pronounced for tests in mathematics and older children. The last four bars show that the results are robust to two

important robustness checks: using only data on the subsample of schools we included in the break survey and using only within students' variation in test hour (i.e., including individual fixed effects). In this individual fixed effects specification, we remove any individual time-and-subject-invariant unobserved effect, but still find the same pattern of improvements during breaks and deterioration for every hour later in the day the test is taken. These effects, therefore, are not driven by selection of students into specific times of the day.

Fig. 2B shows the heterogeneous effects of test hour and breaks on different percentiles of the test score distributions. The graph was created based on quantile regressions and shows the effect of breaks and test hour for different percentiles of the test score distribution. This analysis shows that both breaks and time of day affect the lower end of the distribution, i.e., the low performing students, significantly more than the upper end of the test score distribution, i.e., the high performing students. For the 10th percentile, a break causes 2.7% (95% CI, 2.0–3.5%) of an SD improvement in test score, and for every hour later in the day, the test performance worsens by 1.3% (95% CI, 1.0–1.5%) of an SD. At the upper end of the distribution, there is no effect of breaks on performance, and for every hour, the test score declines by only 0.4% (95% CI, 0.2–0.6%) of an SD.

Overall, the results of our analyses provide support for our hypotheses that taking tests later in the day worsens performance and taking tests after a break improves performance.

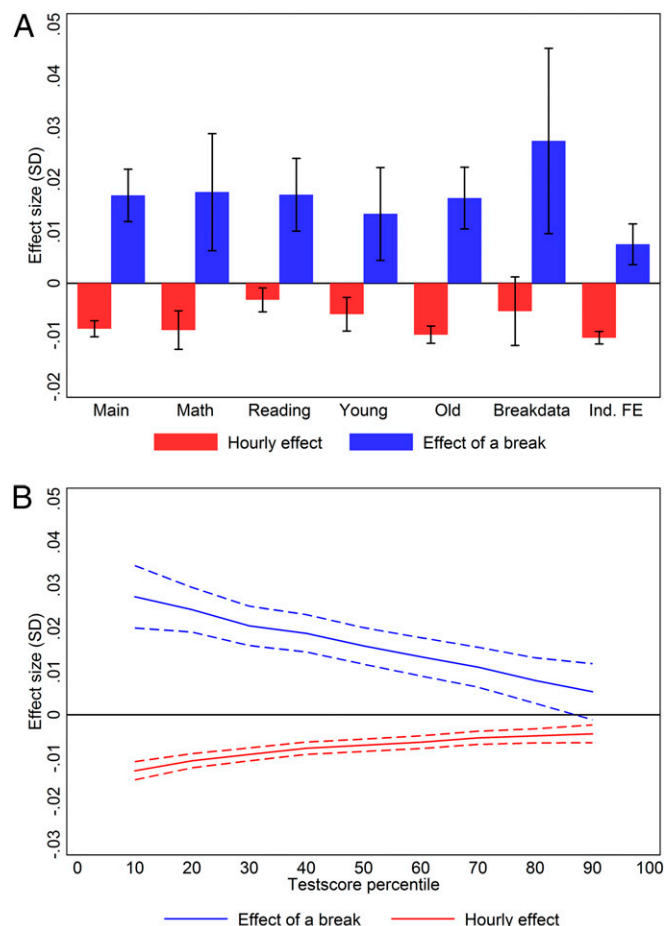
## Conclusion

Standardized testing is commonly used to assess student knowledge across countries and often drives education policy. Despite its implications for students' development and future, it is not without bias. In this paper, we examined the influence of the time at which students take tests and of breaks on test performance. In Denmark, as in many other places across the globe, test time is determined by the weekly class schedule and computer availability at schools. We find that, for every hour later in the day, test scores decrease by 0.9% SDs. In addition, a 20- to 30-minute break improves average test scores. Importantly, a break causes an improvement in test scores that is larger than the hourly deterioration. Therefore, if there was a break after every hour, test scores would actually improve over the day. However, if, like in the Danish system, there is only a break every other hour, the total effect is negative. Our results also show that low-performing students are those who suffer more from fatigue and benefit more from breaks. Thus, having breaks before testing is especially important in schools with students who are struggling and performing at low levels.

To understand effect sizes, we computed the simple correlations between test score and parent income, parental education, and school days (see *SI Text* for details). We find that an hour later in the day causes a deterioration in test score that corresponds to 1,000 USD lower household income, a month less parental education, or 10 school days. A break causes an improvement in test score that corresponds to about 1,900 USD higher household income, almost 2 months of parental education, or 19 school days. The effect sizes are small but nonnegligible compared to the unconditional influence of individual characteristics.

Importantly, the students in our sample are young children and early adolescents, and older adolescents may fare differently. We hope future research will investigate this possibility. Future work could also examine other forms of potential variation in students' performance on standardized tests, including circadian rhythms (22). In fact, research has shown individuals' cognitive functioning (e.g., memory and attention) is at its peak at their optimal time of day and decreases substantially at their nonoptimal times (23–25).

Our results should not be interpreted as evidence that the start time of the school day should change to later (thus allowing students to sleep in, as currently debated in the United States) or



**Fig. 2.** Effect of time of day and breaks. Effects are estimated based on administrative data from Statistics Denmark. The figures show the parameter estimates for break and test hour from estimating a linear model of test score on test hour and break and controlling for test year fixed effects, as well as parental income, parental education, nonwestern origin, sex, spring child, and birth weight. We also control for school, grade, subject, and day of the week fixed effects. The details on the model and estimation procedure are shown in *SI Text*, along with a table with regression results. (A) Main effect and the effect by subgroups. (B) Results from quantile regression at the 10th, 20th, 30th, 40th, ..., 90th percentiles using Canay's plugin fixed effect estimator (21). The graph shows the effect of breaks and test hour over the test score distribution.



that schools tests should be administered earlier in the day. Rather, we believe these results to have two important policy implications: first, cognitive fatigue should be taken into consideration when deciding on the length of the school day and the frequency and duration of breaks. Our results show that longer school days can be justified, if they include an appropriate number of breaks. Second, school accountability systems should control for the influence of external factors on test scores. How can school systems handle such potential biases? One approach would be to adjust the test scores based on the parameters identified in this paper. Based on our results, policy makers should adjust upward test scores by 0.9% of an SD for every hour later in the day the test is taken, and adjust downward tests after breaks with 1.7% of an SD. We recognize that this approach may not always be feasible to implement in practice given that it would require continuous monitoring and adjustments. A more straightforward approach would be to plan tests as closely after breaks as possible. Moreover, as breaks and time of day clearly affect students' test performance, we also expect other external factors like hunger, light conditions, and noise to play a role. These external factors should be accounted for when comparing test scores across children and schools.

## Data and Methods

Here we describe how to obtain access to the data analyzed in our paper. For additional methodological detail, full results, and tables, please refer to [SI Text](#). The project was carried out under Agreement 2015-57-0083 between The Danish Data Protection Agency and the Danish National Centre for Social Research. Specifically, this study was approved by the research board of the Danish National Centre for Social Research under Project US2280 and approved by the Danish protection agency under Agreement 2015-57-0083. We note that there is no Danish institutional review board for studies that are not randomized controlled trials.

The analyses are based on data from administrative registers on the Danish population provided by Statistics Denmark and the Danish Ministry for Education. All analyses have been conducted on a server hosted by Statistics

Denmark and owned by The Danish National Centre for Social Research (SFI server project number 704335). All calculations were done with the software STATA (version 13.0). Given that these data contain personal identifiers and sensitive information for residents, they are confidential under the Danish Administrative Procedures (§27) and the Danish Criminal Code (§152). Therefore, we cannot make the data publicly available. However, independent researchers can apply to Statistics Denmark for access, and we will assist in this process in any way we can. If interested researchers request and obtain access to the data, they can use the stata code included in [SI Appendix](#) to reproduce the results of the analyses reported in the paper and in the [SI Text](#).

Statistics Denmark requires that researchers who access the confidential information receive approval by a Danish Research Institute. The Danish National Centre for Social Research is willing to grant researchers access to this project, given that they satisfy the existing requirements. As of today, the formal requirements involve a test in data policies and a signed agreement. More information on the Danish National Centre for Social Research can be found at [www.sfi.dk](http://www.sfi.dk).

The Danish Ministry for Education granted us access to all test results from the mandatory National Tests in Danish Public Schools between school years 2009/2010 and 2012/2013. The data were sent from the Ministry to Statistics Denmark. Statistics Denmark anonymized the personal identifiers and provided information on each student's birth weight, parental income rank, parental education (years), and sex. Before analyzing the data, we excluded 14,945 tests that were taken at 2:00 PM and 2,918 tests that were taken in grades and subject combinations that are out of schedule. The pattern of results for tests occurring at 2:00 PM is in line with the overall conclusions we draw in our research, but this test time was so uncommon that we excluded it from the sample. In total we excluded 17,863 of 2,052,827 observations (0.9% of the raw sample). All conclusions remain unchanged if we conduct the analyses on the raw sample. The sample selection is done to ensure that the analysis is based on normal tests and not tests that were taken under special circumstances.

**ACKNOWLEDGMENTS.** We thank seminar participants from the Copenhagen Education Network for comments. We appreciate the helpful comments we received from Ulrik Hvidman, Mike Luca, Alessandro Martinello, and Todd Rogers on earlier drafts. H.H.S. acknowledges financial support from Danish Council for Independent Research Grant 09-070295.

1. US Legal I (2014) Standardized test [education] law & legal definition. Available at [definitions.uslegal.com/s/standardized-test-education/](http://definitions.uslegal.com/s/standardized-test-education/). Accessed January 29, 2016.
2. Robelen EW (2002) An ESEA primer. *Educ Week* February:21.
3. Koretz D, Deibert E (1996) Setting standards and interpreting achievement: A cautionary tale from the National Assessment of Educational Progress. *Educ Assess* 3(1): 53–81.
4. Mullette-Gillman OA, Leong RLF, Kurnianingsih YA (2015) Cognitive fatigue destabilizes economic decision making preferences and strategies. *PLoS One* 10(7):e0132022.
5. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB (2001) The job demands-resources model of burnout. *J Appl Psychol* 86(3):499–512.
6. Holding D (1983) *Fatigue. Stress and Fatigue in Human Performance* (John Wiley & Sons, New York).
7. Boksem MA, Meijman TF, Lorist MM (2005) Effects of mental fatigue on attention: An ERP study. *Brain Res Cogn Brain Res* 25(1):107–116.
8. Lorist MM, Boksem MA, Ridderinkhof KR (2005) Impaired cognitive control and reduced cingulate activity during mental fatigue. *Brain Res Cogn Brain Res* 24(2):199–205.
9. Sanders AF (1998) *Elements of Human Performance: Reaction Processes and Attention in Human Skill* (Lawrence Erlbaum Associates, London).
10. van der Linden D, Frese M, Meijman TF (2003) Mental fatigue and the control of cognitive processes: Effects on perseveration and planning. *Acta Psychol (Amst)* 113(1):45–65.
11. Boksem MA, Meijman TF, Lorist MM (2006) Mental fatigue, motivation and action monitoring. *Biol Psychol* 72(2):123–132.
12. Hockey GRJ, John Maule A, Clough PJ, Bdzola L (2000) Effects of negative mood states on risk in everyday decision making. *Cogn Emotion* 14(6):823–855.
13. Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. *Proc Natl Acad Sci USA* 108(17):6889–6892.
14. Vohs KD, et al. (2008) Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *J Pers Soc Psychol* 94(5):883–898.
15. Linder JA, et al. (2014) Time of day and the decision to prescribe antibiotics. *JAMA Intern Med* 174(12):2029–2031.
16. Henning RA, Sauter SL, Salvendy G, Krieg EF, Jr (1989) Microbreak length, performance, and stress in a data entry task. *Ergonomics* 32(7):855–864.
17. Gilboa S, Shirom A, Fried Y, Cooper C (2008) A meta-analysis of work demand stressors and job performance: Examining main and moderating effects. *Person Psychol* 61(2): 227–271.
18. Smith SM (1995) Getting into and out of mental ruts: A theory of fixation, incubation, and insight. *The Nature of Insight*, eds Sternberg RJ, Davidson JE (MIT Press, Cambridge, MA), pp 229–251.
19. Leonard D, Swap W (1999) *When Sparks Fly: Igniting Creativity in Groups* (Harvard Business School Press, Boston).
20. Schön DA (1983) *The Reflective Practitioner: How Professionals Think in Action* (Basic Books, New York).
21. Canay IA (2011) A simple approach to quantile regression for panel data. *Econ J* 14(3): 368–386.
22. Yoon C, May CP, Hasher L (1999) Aging, circadian arousal patterns, and cognition. *Aging, Cognition and Self Reports*, eds Schwarz N, Park D, Knauper B, Sudman S (Psychological Press, Washington, DC), pp 117–143.
23. Blake MJF (1967) Time of day effects on performance in a range of tasks. *Psychon Sci* 9(6):349–350.
24. Goldstein D, Hahn CS, Hasher L, Wiprzycka UJ, Zelazo PD (2007) Time of day, intellectual performance, and behavioral problems in morning versus evening type adolescents: Is there a synchrony effect? *Pers Individ Dif* 42(3):431–440.
25. Randler C, Frech D (2009) Young people's time-of-day preferences affect their school performance. *J Youth Stud* 12(6):653–667.
26. Hayashi F (2000) *Econometrics* (Princeton Univ, Princeton).
27. Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. *J Hum Resour* 50(2):317–372.

# Supporting Information

Sievertsen et al. 10.1073/pnas.1516947113

## SI Text

**Data.** The analyses are based on data from administrative registers on the Danish population provided by Statistics Denmark and the Danish Ministry for Education. The Danish Ministry for Education granted us access to all test results from the mandatory National Tests in Danish Public Schools between school years 2009/2010 and 2012/2013. The data were sent from the Ministry to Statistics Denmark. Statistics Denmark anonymized the personal identifiers and provided information on each student's birth weight, parental income, parental education (years), and sex. Before analyzing the data, we excluded 14,945 tests that were taken at 2:00 PM and 2,918 tests that were taken in grades and subject combinations that are out of schedule. The pattern of results for tests occurring at 2:00 PM is in line with the overall conclusions we draw in our research, but this test time was so uncommon that we excluded it from the sample. In total we excluded 17,863 of 2,052,827 observations (0.9% of the raw sample). All conclusions remain unchanged if we conduct the analyses on the raw sample. The sample selection is done to ensure that the analysis is based on normal tests and not tests that were taken under special circumstances.

**Methodology.** The test data contain information on each individual's test time and test performance. The main analysis is based on a comparison of mean test scores over test times. This mean comparison provides an estimate of the causal effect of test time on test performance given that the test time is not correlated with any observed or unobserved individual characteristics. In the Danish setting, variation in test time is created by the way the tests are planned. Specifically, a test time is selected based on three criteria: (i) the teacher has to plan the test in the weekly schedule of the given subject; (ii) the computer facilities must be available; and (iii) the test must be taken in the Spring term. The test time is then determined according to these criteria, and often the availability of computer facilities is the binding constraint. We claim that this creates variation in test time that is as good as random to the individual. We provide support for this claim later in SI Text.

To test our hypothesis regarding the effect of time of the test on test score, we first estimate a linear relationship between test score and test hour

$$\text{testscore}_{it} = \alpha_0 + \alpha_1 \text{testhour}_{it} + \epsilon_{it}. \quad [\text{S1}]$$

Model S1 is estimated using OLS. It estimates the correlation between the time of the day when the test is taken and the test score. However, as breaks are likely to influence students' fatigue, we relax the linear assumption and allow each test hour to have a separate effect. In our preferred specification, we exploit variation within schools (i.e., school fixed effects), within grades (i.e., grade fixed effects), within days of the week (i.e., day of the week fixed effects), within test years (i.e., test year fixed effects), and within subjects (i.e., subject fixed effect). The estimation equation for Fig. 1 is thus given by

$$\begin{aligned} \text{testscore}_{it} = & \alpha_0 + \alpha_1 9_{it} + \alpha_2 10_{it} + \alpha_3 11_{it} + \alpha_4 12_{it} + \alpha_5 13_{it} \\ & + \tau \text{DOW}_{it} + \beta \text{Year}_{it} + \gamma \text{Subject}_{it} + \theta \text{Grade}_{it} \\ & + \delta \text{School}_{it} + \mu X_{it} + \epsilon_{it}, \end{aligned} \quad [\text{S2}]$$

where 9–13 are indicators for the hour the test  $t$  of individual  $i$  started. *Year*, *Subject*, *DOW*, *Grade*, and *School* are vectors of indicators for the test year, subject, day of the week, grade, and school,

and they thus capture the fixed effects.  $X$  is a vector of individual control variables: parental education, household income, household income rank, birth weight, spring birth, sex, and nonwestern origin. Covariates are measured in the calendar year before the test is taken. Model S2 is estimated using OLS.

**Distributional assumptions.** As the dataset we use in our analyses includes more than two million observations, we can rely on the asymptotic properties (see chapter 2 in ref. 26). However, we assess the distribution of the dependent variable in the following sections to fully understand the nature of our outcome variable of interest. **SEs.** As both observed and unobserved factors might be correlated within schools, we allow arbitrary correlation within these units when computing the SEs, according to the recommendations in ref. 27.

**Missing values.** For a small fraction of the sample, we are not able to match individual background characteristics to the test score data. This attrition is balanced across test hours. In the regressions, we include indicators for missing values and replace the missing value with a zero. This strategy is only valid if the values are missing at random. Missing values most likely occur because the student was not living in Denmark the year before the test. Results only change slightly when covariates are excluded, which suggests that covariates are not systematically correlated with test hour and missing values are not systematically correlated with test hour. Despite the minimal change in results we obtain when we exclude covariates, the significance of the results is not affected.

**Discussion of main results.** The upper part of Fig. 1 shows the hour-to-hour change in test scores. The point-estimate for coefficient  $\alpha_1$  gives the estimate for the change in test score when the test is at 9:00 AM instead of 8:00 AM. The coefficient  $\alpha_2$  gives the point-estimate for the change in test score when the test is at 10:00 AM instead of 8:00 AM. We obtain the point estimate for the change in test score from 9:00 AM to 10:00 AM from  $\alpha_2 - \alpha_1$ , which gives the height of the bar for 10:00 AM in Fig. 1. The SE for this difference is calculated using the estimated covariance matrix:  $se(\hat{\alpha}_2 - \hat{\alpha}_1) = \sqrt{\text{var}(\hat{\alpha}_1) + \text{var}(\hat{\alpha}_2) - 2\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2)}$ . The error bars in Fig. 1 are calculated by  $(\hat{\alpha}_2 - \hat{\alpha}_1) \pm 1.96 * se(\hat{\alpha}_2 - \hat{\alpha}_1)$ .

We used this specification because it allows every test hour to have its own effect. As the upper part of Fig. 1 shows, a test taken at 9:00 AM instead of 8:00 AM causes a test score reduction of 0.35% of an SD. Likewise, the bar at 10:00 AM, with a height of 0.07% of an SD, shows that test scores improve by 0.07% of an SD from 9:00 AM to 10:00 AM (the height of this bar is computed by  $0.35 - 0.28$ , based on the estimates at 9 and 10, in column 4 in Table S1).

Next, we focus on the effect of having the test after a typical break. By typical break, we mean a break that commonly occurs at the same time throughout the week, across schools. The lower part of Fig. 1 shows the distribution of breaks across schools. Specifically, it plots the time when the breaks commonly end. We focus on the time when the breaks end because breaks vary in their length. We contacted 10% of the 2,105 schools in our dataset and asked about the break schedule. We received responses from 95 schools, corresponding to a 45% response rate. Eighty-four percent of the schools we interviewed confirmed they follow the usual break schedule on test days. As the lower part of Fig. 1 shows, almost all schools have breaks before the 10:00 AM test and again before the 12:00 PM test.

Because we hypothesized that breaks cause test improvements, we also estimate a more restricted version of the relationship between test hour and test performance, where we include a variable for test hour (as in model S1) and an indicator for whether the test

was taken after a typical break. This approach is used for Fig. 2. The estimation equation for Fig. 2 is thus given by

$$\text{testscore}_{it} = \alpha_0 + \alpha_1 \text{break}_{it} + \alpha_2 \text{testhour}_{it} + \mu X_{it} + \tau \text{DOW}_{it} + \beta \text{Year}_{it} + \gamma \text{Subject}_{it} + \theta \text{Grade}_{it} + \delta \text{School}_{it} + \epsilon_{it}, \quad [\text{S3}]$$

where break is an indicator for the test being taken at a time that is usually after a break (8:00 AM, 10:00 AM, 12:00 PM). Fig. 24 shows the point estimates and error bars for the coefficients  $\alpha_1$  and  $\alpha_2$ . The bar “Main” is for the full sample. The bar “Math” only includes tests in Mathematics. The bar “Reading” only includes tests in Reading/Danish. The bar “Young” only includes tests taken in grades 2–4. The bar “Old” only includes tests taken in grades 6–8. The bar “Break data” only includes tests taken in the schools with information on the break schedule. The bar “Individual FE” is for a version where we include individual fixed effects and thus only exploit variation in test time from year to year or subject to subject, within the same student. As the upper part of Fig. 24 shows, test scores worsen for every hour later in the day the test is taken (the red bars), but improve after a break (the blue bars). This pattern holds across subsamples and specifications.

Fig. 2B shows the results from a quantile regression of model S3 on the 10th, 20th, ..., 90th percentiles using the Canay plug-in fixed effects approach (22). The OLS estimation above estimates the effect on the mean. With a quantile regression we can estimate the effect on specific quantiles (or percentiles) of the test distribution. Although the estimation used above shows how average test scores change after breaks and during the day, this regression shows how the median (or for example the 10th percentile) test scores are affected by time of the day. In a linear regression using OLS, it is straightforward to control for the school fixed effects, as the linear specification allows us to cancel these effects out. The same approach cannot be used with quantile regressions, as the estimation is nonlinear. Canay (22) introduced an estimator where we first clean the dependent variable for the fixed effects (i.e., the unobserved school fixed effects), and then use this cleaned dependent variable in the quantile regression framework, which is the methodology we apply here.

Table S1 shows the results from estimating the main models. Column 1 shows the estimates from estimating model S1. The first row shows that for every hour later in the day, the test score is reduced by 0.6% of an SD. In column 2, we relax the assumption of linearity and allow each test hour to have its own effect. The point estimate in row 9 shows that tests at 9:00 AM on average are 5% of an SD worse than tests at 8:00 AM. The point estimate in row 10 shows that tests at 10:00 AM on average are 3.8% of an SD worse than tests at 8:00 AM, and thus 1.2% points better than tests at 9:00 AM. Tests at 11:00 AM are 6% of an SD worse than at 8:00 AM and 2.2% points worse than at 10:00 AM. The point estimates in column 2 of row 12 show that at 12:00 PM test scores are on average 3.3% of an SD lower than at 8:00 AM, and finally row 13 shows that tests at 1:00 PM are 3.9% of an SD worse than tests at 8:00 AM. Together, these results show that taking the test earlier in the day is advantageous to students as it is related to higher test scores. This pattern of results is likely the result of increased cognitive fatigue as the day wears on, consistent with our theorizing based on existing work in psychology.

In column 3 of Table S1, we further added school, test year, grade, subject, and day of the week fixed effects. In this specification, we only include schools with at least two tests (which forces us to drop 77 schools/observations out of 2,034,965 observations). Although in column 2 we show how average test scores change from hour to hour, overall, in column 3 we show how test scores change from hour to hour, holding the school, day of the week, test year, subject, and grade constant. Controlling for these factors reduces the magnitude of the estimates somewhat, but the overall pattern

remains unchanged. In column 4, we further added individual controls, namely child birth weight, parental education, parental income, sex, and origin. Adding these covariates only has a minor effect on the point estimates, and the conclusion remains unchanged.

In column 5 of Table S1, we show point estimates from estimating model S3. The first row shows that for every hour later in the day, test scores worsen by 0.5% of an SD, but the estimates for break in the next row shows that, after breaks, test scores improve by 2.5% of an SD. Although column 5, like column 2, is a simple comparison across all schools and subjects, column 6, just like column 3, also controls for fixed effects and thus shows the effect of breaks and test hour while holding other factors fixed. This change increases the effect of time of day on test scores slightly, from 0.5% to 0.8% of an SD, and reduces the effect of breaks from 2.5% to 1.9% of an SD. In column 7, we further added individual controls, which only causes a negligible effect on the point estimates. Once we keep both fixed effects and individual characteristics constant, we find that the test scores worsen by 0.9% of an SD for every hour later in the day the test is taken, and breaks cause an improvement of 1.7% of an SD.

The lower part of Table S1 shows various model characteristics and diagnostics. We include 2,028 schools in the sample. For the smallest school, we only have two tests, whereas for the largest school, we have 3,984 observations. Next we conduct an *F*-test of the joint significance for the hourly indicators in columns 2–4. They all clearly show that we reject the null hypotheses that these coefficients are zero. The adjusted  $R^2$  is very low across specifications, which is not uncommon in microeconometrics. In columns 4 and 7, which are the preferred specifications used to create Figs. 1 and 2, the adjusted  $R^2$  is 0.08.

**Distribution of Test Scores.** As mentioned above, we rely on asymptotic characteristics of the estimator for both the estimation of the point-estimates and the inference. However, to better understand the data and the results, it is useful to consider the distribution of the test score (our main dependent measure). Fig. S1 shows kernel density estimates for the raw test scores, by test hour. The distributions seem normal, but with somewhat long tails. The graph shows that there are no systematic differences in the overall distribution of tests across test times, with the exception of slight shifts of the distributions along the *x* axis. In other words, tests at 9:00 AM are shifted slightly to the right compared with tests at 8:00 AM, indicating worse test performance at 9:00 AM than at 8:00 AM, but the overall shape of the test score distribution does not differ across the distribution (e.g., the shape of the distribution at 9:00 AM looks similar to the distribution at 11:00 AM). The results of this analysis give us confidence that our results are not driven by obscure properties of the distribution of test scores. In addition, this gives us confidence that the regression analyses we conducted to examine the effects of time of day and breaks on test scores are valid, and are not subject to bias.

**As Good as Random Variation in Test Time.** As mentioned earlier, the results we obtained in the main analyses we conducted can be interpreted as causal effects of test hour/breaks on test performance if, and only if, the test time is as good as random. We assess this assumption in three ways:

- i) We regress the covariates (birth weight, parental income, parental education) on test hour and compare the *z*-scores (*t*-values) across sample sizes with the *z*-scores from the main analyses. The intuition is that, although for specifications with test scores as dependent variables the hourly indicators are expected to be precisely estimated, the specifications using covariates as dependent variables should not give very precise estimates of the hourly indicators, as this would indicate that the hourly indicators are closely correlated with individual characteristics. The plotted *z*-scores are used to compute *P* values, and larger *z*-scores will result in lower *P* values.









**Table S1. Main regression results**

Variable	1	2	3	4	5	6	7
<i>testhour</i>	−0.006*** (0.001)						
<i>break</i>							
9		−0.050*** (0.009)	−0.038*** (0.004)	−0.035*** (0.004)	−0.005*** (0.001)	−0.008*** (0.001)	−0.009*** (0.001)
10		−0.038*** (0.005)	−0.029*** (0.003)	−0.028*** (0.003)	0.025*** (0.006)	0.019*** (0.003)	0.017*** (0.003)
11		−0.060*** (0.009)	−0.055*** (0.004)	−0.052*** (0.004)			
12		−0.033*** (0.007)	−0.042*** (0.004)	−0.042*** (0.004)			
13		−0.039*** (0.011)	−0.052*** (0.006)	−0.054*** (0.006)			
Fixed effects included	No	No	Yes	Yes	No	Yes	Yes
Individual controls	No	No	No	Yes	No	No	Yes
Number of schools			2,028	2,028		2,028	2,028
Smallest group			2	2		2	2
Largest group			3,984	3,984		3,984	3,984
$F$ -value $th9 = th10$ := $th13 = 0$		14.84	38.61	39.59			
$P$ value $th9 = th10$ := $th13 = 0$		0.00	0.00	0.00			
Model degrees of freedom		5	20	30	2	17	27
Adjusted $R^2$		0.00	0.00	0.08	0.00	0.00	0.08
AIC	5,774,756	5,774,188	5,626,905	5,462,886	5,774,577	5,627,291	5,463,303
Observations	2,034,964	2,034,964	2,034,887	2,034,887	2,034,964	2,034,887	2,034,887

The dependent variable in each model is standardized test score. All estimates are obtained using ordinary least squares (OLS). Column 1 shows the point estimate for  $\alpha_1$  from estimating model S1. Columns 2–4 show results from estimating model S2. Columns 5–7 show results from estimating model S3. The table only shows the point estimates for the coefficients  $\alpha_1 - \alpha_5$  in model S2 and coefficients  $\alpha_1$  and  $\alpha_2$  for model S3. Columns 2 and 5 show results from estimating models without any control variables. Columns 3 and 6 show results from estimating simple models with only school, year, day of the week, grade, and subject fixed effects. Columns 4 and 7 show results from estimating the full models without individual fixed effects. SEs clustered at the school level are shown in parentheses. Number of schools show the number of schools included and thus also the level of fixed effects and clustering. Smallest/largest group shows the smallest/largest number of observations from one school.  $F$ -value gives the  $F$ -statistic for a test of joint significance for the hourly indicators in columns 2–5, and  $P$  value gives the corresponding  $P$  values. This  $P$  value is for the null hypothesis that all hourly indicators are zero. Model degrees of freedom specifies the number the degrees of freedom used by the model. AIC gives the Akaike information criteria. Smaller AICs are generally preferred. The term observations refers to the number of observations included in the regressions. The dependent variable is standardized within the test year, grade, and subject cell. As fixed effects on the school level implies comparisons within schools, we only include schools with at least two tests. Regressions are based on administrative data from Statistics Denmark and the Danish Ministry for Education, for all mandatory tests 2009/10–2012/13.

\*\*\* $P < 0.01$ ; \*\* $P < 0.05$ ; \* $P < 0.1$ .

**Table S2. Variable means across test time**

		Test hour					
Variable	All	8	9	10	11	12	13
Full sample							
Test uncertainty	0.26	0.26	0.27	0.26	0.28	0.26	0.26
School day	179.89	180.10	180.02	178.81	180.49	180.83	179.80
Child birth weight (g)	3,298	3,307	3,304	3,300	3,303	3,282	3,272
Parents' years of schooling	14.24	14.26	14.18	14.21	14.25	14.30	14.33
Household income, 1,000 DKK	386.17	384.64	379.79	386.07	382.60	396.16	396.96
Household income percentile	57.05	57.14	56.32	56.98	56.77	57.91	58.16
Nonwestern origin	0.09	0.08	0.09	0.09	0.08	0.09	0.10
Female	0.49	0.49	0.48	0.49	0.49	0.49	0.49
Spring child	0.50	0.50	0.49	0.49	0.50	0.50	0.50
Missing birth weight data	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Missing education data	0.03	0.03	0.03	0.03	0.02	0.02	0.01
Observations	2,034,964	400,139	424,868	528,764	259,649	295,032	126,512
Subsample of schools included in the break data sample							
Test uncertainty	0.24	0.24	0.22	0.25	0.27	0.24	0.25
School day	179.65	179.18	179.57	180.18	179.61	179.86	178.93
Child birth weight (g)	3,305	3,302	3,303	3,303	3,321	3,295	3,323
Parents' years of schooling	14.29	14.26	14.17	14.30	14.35	14.38	14.40
Household income, 1,000 DKK	389.37	392.55	376.89	393.87	385.21	397.78	390.23
Household income percentile	57.25	57.43	56.11	57.42	57.06	58.10	58.18
Nonwestern origin	0.08	0.08	0.08	0.09	0.06	0.08	0.08
Female	0.49	0.50	0.49	0.49	0.49	0.49	0.48
Spring child	0.50	0.49	0.50	0.49	0.49	0.50	0.51
Missing birth weight data	0.06	0.06	0.06	0.06	0.05	0.06	0.05
Missing education data	0.03	0.03	0.03	0.03	0.02	0.02	0.01
Observations	121,709	25,561	24,025	31,198	17,039	16,995	6,891

The table shows simple averages for subsamples across test hours. Test uncertainty is the estimated uncertainty for the test result (only available for years 2011–2013). This variable measures how precisely the individual test score was estimated by the computer. (The uncertainty comes from the fact that the test can be more or less accurate to measure the student performance. Each question contributes to the precision of the test, reducing uncertainty. That is, the test is adaptive, and the computer tries to calculate the individuals level. For instance, if the student reaches a level where every time he/she gets a harder question, he/she answers wrong, but every time he/she gets an easier question, he/she answers right, there will be very low uncertainty on test score.) Income is adjusted to the 2010 price level and adjusted for household size using the square root approach. Spring child is a child born in the period January–June of the year being considered. Because the school starting age cutoff is January 1, these children have the highest expected school starting age. The variable called school days refers to the number of days from the start of the school year to the test day, not counting weekends. Parents' years of schooling is the years of schooling completed by the mother or father (the highest value).







**Table S5. Regression results**

Variable	1	2	3	4
Household income	0.0013*** (0.0000)			
Birth weight		0.0001*** (0.0000)		
Years of schooling			0.1195*** (0.0010)	
School days				0.0009*** (0.0001)
Adjusted $R^2$	0.06	0.00	0.10	0.00
AIC	5,554,828	5,361,776	5,404,795	5,773,793
$N$	2,002,033	1,899,159	1,981,671	2,034,964

Dependent variable: standardized test score. SEs clustered at the school level are shown in parentheses. AIC gives the Akaike information criteria. Observations shows number of observations included in the regressions. The dependent variable is standardized within the test year, grade, and subject cell. We removed outliers in income and birth weight (first and last percentile). This is done because, in this linear regression, measurement errors (e.g., birth weight of 10 kg) would have a huge impact on the point estimates. However, overall the conclusions are not very sensitive to this change.

\*\*\* $P < 0.01$ ; \*\* $P < 0.05$ ; \* $P < 0.1$ .

## Other Supporting Information Files

[SI Appendix \(PDF\)](#)



```
1  * time of day and test performance
2  * Last edited: 20160113 by hhs@sfi.dk
3  ****
4  clear all
5  set max_memory 2g, perm
6
7  * Set directory
8  cd "D:\Data\workdata\704335\Timeofday"
9
10 global tf "D:\Data\workdata\704335\Timeofday\tempfiles"
11 global df "D:\Data\workdata\704335\Timeofday\download"
12 global rf "D:\Data\workdata\704335\stataraw_new"
13
14 adopath + "D:\Data\workdata\704335\Timeofday\adofiles"
15
16 * Controls
17 global covariates "birthweight edu inc incrank immigr_nonwestern female spring "
18 global missings "missing_birthweight missing_edu missing_inc missing_incrank
19 missing_immigr_nonwestern missing_female missing_spring"
20 global controls1 "i.sub i.grade i.dow i.testyear"
21 global controls2 "$covariates $missings $controls1"
```

```

1
2  /*****
3  * Create data for time of day project
4  * Last edited: 20160113 by hhs@sfi.dk
5  *****/
6
7  * Load preamble
8  do "D:\Data\workdata\704335\Timeofday\dofiles\preamble.do"
9  * Set memory
10 set max_memory 12g, perm
11
12 /*****
13 * Outline:
14 * 1: Create covariate data
15 * 2: Create test data
16 * 3: Merge 1 and 2
17 *****/
18
19
20
21
22 /*****
23 * 1 Create covariate data
24 *****/
25 * Load "Grund" data for 1995 to 2012 and make sure everything is in the right format.
26 forval i=1995/2012{
27     use "$rf\GRUND`i'.dta", clear
28     * foed_dag changes name after 2006. Correct this
29     if `i'>2006{
30         rename FOED_DAG foed_dag
31     }
32     * Keep what we need
33     keep pnr familie_id koen kom IE_TYPE OPR_LAND hfaudd SAMLINK_NY foed_dag
34     * Education level within the family
35     rename hfaudd audd
36     tostring audd, replace
37     merge m:1 audd using "$tf\audd.dta"
38     drop if _merge==2
39     drop _merge
40     bys familie_id: egen educ=max(pria)
41     gen edu=educ/12
42
43     * Household income
44     * Adjust to 2010 level
45     gen gross_inc=SAMLINK_NY
46     replace gross_inc=gross_inc*(122.4/89.2) if `i'==1995
47     replace gross_inc=gross_inc*(122.4/91.1) if `i'==1996
48     replace gross_inc=gross_inc*(122.4/93.1) if `i'==1997
49     replace gross_inc=gross_inc*(122.4/94.8) if `i'==1998
50     replace gross_inc=gross_inc*(122.4/97.2) if `i'==1999
51     replace gross_inc=gross_inc*(122.4/100.0) if `i'==2000
52     replace gross_inc=gross_inc*(122.4/102.4) if `i'==2001
53     replace gross_inc=gross_inc*(122.4/104.8) if `i'==2002
54     replace gross_inc=gross_inc*(122.4/107.0) if `i'==2003
55     replace gross_inc=gross_inc*(122.4/108.3) if `i'==2004
56     replace gross_inc=gross_inc*(122.4/110.2) if `i'==2005
57     replace gross_inc=gross_inc*(122.4/112.3) if `i'==2006
58     replace gross_inc=gross_inc*(122.4/114.2) if `i'==2007
59     replace gross_inc=gross_inc*(122.4/118.1) if `i'==2008
60     replace gross_inc=gross_inc*(122.4/119.7) if `i'==2009
61     replace gross_inc=gross_inc*(122.4/122.4) if `i'==2010
62     replace gross_inc=gross_inc*(122.4/125.8) if `i'==2011
63     replace gross_inc=gross_inc*(122.4/128.8) if `i'==2012
64     * Total income on household level
65     bys familie_id: egen inc=sum(gross_inc)
66     * Family members
67     bys familie_id: gen count=_n
68     bys familie_id: egen members=max(count)
69     * Adjusted household income
70     replace inc=inc/(members^0.5)
71     replace inc=inc/1000
72     * Household income rank
73     gen incl=inc if count==1
74     egen income=xtile(incl),nq(100)
75     bys familie_id: egen incrank=min(income)

```

```

76
77     * Country formats
78     rename OPR_LAND land
79     merge m:1 land using "$tf\ieland.dta"
80     drop if _merge==2
81     gen western=VEST_EJ==1|VEST_EJ==2
82     gen immigr_nonwestern=(IE_TYPE==2|IE_TYPE==3)&western==0
83     * Gender
84     gen female=koen-1
85     * Spring child
86     gen spring=month(foed_dag)<7
87     * Variable labels
88     label var spring "Spring child"
89     label var inc "Household income, 1,000 DKK"
90     label var immigr_nonwestern "Nonwestern immigrant/desc."
91     label var female "Female"
92     label var incrank "Household inc. percentile"
93     label var edu "Parents' years of schooling"
94     * Keep what we need
95     keep kom pnr western immigr_nonwestern female incrank edu inc spring
96
97     gen y=`i'
98     compress
99     save "$tf\GRUND`i'.dta", replace
100 }
101
102 *Append to one dataset
103 clear
104 use "$tf\GRUND1995.dta", clear
105 forval i=1996/2012{
106     append using "$tf\GRUND`i'.dta",
107     }
108
109 save "$tf\covariates.dta",replace
110
111
112 /*****
113 * 2 Create test data
114 *****/
115 * Load
116 use "$rf\DNT2010_2014_HHS_SFI.dta", clear
117 * grade
118 rename klassetrin grade
119 * subject
120 gen subject=substr(fag,1,5)
121 * test time
122 gen testyear=substr(testtid,1,4)
123 gen testmonth=substr(testtid,6,2)
124 gen testday=substr(testtid,9,2)
125 gen testhour=substr(testtid,12,2)
126 gen testmin=substr(testtid,15,2)
127 destring testyear testmonth testday testhour testmin ,replace
128 drop testtid
129 gen date=mdy(testmonth,testday,testyear)
130 * create one uncertainty measure (simple average of the the three)
131 gen uncert=(sem_p1+sem_p2+sem_p3)/3
132 * create one testscore per test
133 gen testscore=(theta_p1+theta_p2+theta_p3)/3
134 bys testyear grade subject: egen sdscore=sd(testscore)
135 bys testyear grade subject: egen mscore=mean(testscore)
136 gen testscore_std=(testscore-mscore)/sdscore
137 * Percentile scores
138 bys testyear grade sub: egen percentile=xtile(testscore), nq(100)
139 label var percentile "Percentile score"
140
141 * create variables and labels
142 * Test hour indicators
143 tab testhour, gen(th)
144 label var th1 "8AM"
145 label var th2 "9AM"
146 label var th3 "10AM"
147 label var th4 "11AM"
148 label var th5 "12Noon"
149 label var th6 "1PM"
150 label var th7 "2PM"

```



```

151 label var testhour "Hour of the day"
152 * Break variables
153 gen break=th1==1|th3==1|th5==1|th7==1
154 gen nobreak=th2==1|th4==1|th6==1
155 label var break "After a break"
156 label var nobreak "Not after a break"
157 * Subject value labels and indicators
158 gen sub=1 if subject=="Dansk"
159 replace sub=2 if subject=="Biolo"
160 replace sub=3 if subject=="Engel"
161 replace sub=4 if subject=="Fysik"
162 replace sub=5 if subject=="Geogr"
163 replace sub=6 if subject=="Matem"
164 label define subl 1 "Danish" 2 "Biology" 3 "English" 4 "Physics" 5 "Geography" 6 "Math"
165 label values sub subl
166 * Day of the week
167 gen dow=dow(date)
168 * Various labels
169 label var uncert "Uncertainty"
170 label var testscore "Testscore (1-100)"
171 label var testscore_std "Standardized testscore"
172 label var grade "Grade"
173 label var dow "Day of the week"
174 label var sub "Subject"
175 * keep, compress and save
176 keep pnr break nobreak testmin uncert instnr percentile date testyear grade sub testyear
177 compress
178 save "$tf\testscoredata.dta",replace
179
180
181
182 /*****
183 * 3 Merge test score data and covariate data
184 *****/
185 use "$tf\testscoredata.dta",clear
186 * merge to birthweight data
187 merge m:1 pnr using "$rf\NYLFOED2010.dta"
188 rename V_VAGT V_VAGT1
189 drop if _merge==2
190 drop _merge
191 merge m:1 pnr using "$rf\MFR2010.dta"
192 drop if _merge==2
193 drop _merge
194 gen birthweight=V_VAGT
195 replace birthweight=V_VAGT1 if V_VAGT==.
196 drop V_GA_DAGE V_APGAR fodtdato V_VAGT1 V_VAGT
197 * merge to covariate panel using the year before the test
198 gen y=year(date)-1
199 merge m:1 pnr y using "$tf\covariates.dta",
200 drop if _merge==2
201 drop _merge
202 * Indicator for break break data
203 gen breakdata=0
204 gen deviates=.
205 destring instnr,replace
206 foreach l in xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx xxxxxx .... xxxxxx {
207 qui: replace breakdata=1 if instnr=="`l'"
208 qui: replace deviates=0 if instnr=="`l'"
209 }
210 foreach l in xxxxxx xxxxxx.... xxxxxx {
211 qui: replace deviates=1 if instnr=="`l'"
212 }
213 * Missing variable indicators
214 foreach l in $covariates {
215 gen missing_`l'=0
216 replace missing_`l'=1 if `l'==.
217 replace `l'=0 if missing_`l'==1
218 }
219 label var missing_birthweight "Missing birthweight data"
220 label var missing_edu "Missing education data"
221 label var missing_immig "Missing origin data"
222 label var missing_female "Missing gender data"
223 label var missing_spring "Missing date of birth"
224 label var missing_inc "Missing income data"

```

```

225 label var missing_incrank "Missing income data"
226 * School day (approximated)
227 gen schoolday=date-mdy(8,1,year(date)-1)
228 * adjust for weekends
229 replace schoolday=schoolday-floor(schoolday/7)*2
230 label var schoolday "School day"
231
232 * PNR as numeric
233 egen id=group(pnr)
234
235 compress
236
237 * Sample selection
238 preserve
239 gen day=testhour+testmin/60
240 keep if testyear==2014
241 drop if day>14
242 hist day, xtitle(Time of the Day) title(Distribution of test times in 2014)
243 graph export "$df\fig2014_testtimedist.pdf",replace
244 tw (lpolyci testscore_std day,bwidth(0.2)), xline(9.5) xline(11.5)
245 graph export "$df\fig2014_testtimeperf.pdf",replace
246 restore
247 drop if testyear==2014
248 qui: sum testscore
249 local nu=r(N)
250 drop if grade==.
251 * tests in nonscheduled grades
252 drop if grade==0
253 drop if grade==1
254 drop if grade==5
255 drop if grade>8
256 drop if grade<8 & sub==2
257 drop if grade==2 & sub==6
258 drop if grade==4 & sub==6
259 drop if grade==7 & sub==6
260 drop if grade==8 & sub==6
261 drop if grade==3 & sub==1
262 drop if grade==7 & sub==1
263 drop if grade!=7 & sub==3
264 drop if grade!=8 & sub==4
265 drop if grade!=8 & sub==5
266 drop if testhour==14
267 sum testscore,d
268 local r1=`nu'-r(N)
269 di "Sampleselection: Deleted `r1' out of `nu' observations"
270 * Keep what we need
271 keep id percentile instnr uncert schoolday break nobreak date testhour testscore
testscore_std th1-th7 testyear sub breakdata deviates grade dow $covariates $missings
272 * final labels
273 label var breakdata "School included in break survey"
274 label var deviates "School deviates from normal schedule when testing"
275 label var birthweight "Child birth weight"
276 label var date "Test date"
277
278 save "$tf\analysisdata.dta",replace
279 keep if runiform()<.01
280 save "$tf\analysisdata1pct.dta",replace
281

```

```

1  /*****
2  * Analyses for test time project, main document
3  * Last edited: 20160113 by hhs@sfi.dk
4  *****/
5
6  * Load preamble
7  do "D:\Data\workdata\704335\Timeofday\dofiles\preamble.do"
8  * Set memory
9  set max_memory 4g, perm
10
11 /*****
12 * Outline:
13 * 1 Main figure hourly bars
14 * 2 The effect of breaks
15 * 3 figure break bars and quantile regs
16 *****/
17
18 /*****
19     1 Main figure hourly bars
20 *****/
21 * Load data
22 use "$tf\analysisdata\sample'.dta",clear
23 * set fixed effect level
24 xtset instnr
25 * singletons in instnr?
26 bys instnr: gen count=_n
27 bys instnr: egen instcount=max(count)
28 * estimate with basic vars
29 eststo: qui: xtreg testscore_std th2-th6 $controls2 if instcount>1, cluster(instnr) fe
30 * create dataset to save estimates
31 clear
32 set obs 5
33 gen th=_n+1
34 * save estimates for testhour==9
35 gen u= _b[th2]+invttail(e(df_r),0.025)*_se[th2]
36 gen l= _b[th2]-invttail(e(df_r),0.025)*_se[th2]
37 gen beta=_b[th2]
38 * save estimates for testhour>9
39 forval i=2/5{
40     local j=`i'+1
41     replace beta= _b[th`j']-_b[th`i'] if th==`j'
42     test th`j'-th`i'=0
43     replace u= _b[th`j']-_b[th`i']+invttail(e(df_r),0.025)*((_b[th`j']-_b[th`i'])/(r(F)^.5))
44     if th==`j'
45     replace l= _b[th`j']-_b[th`i']-invttail(e(df_r),0.025)*((_b[th`j']-_b[th`i'])/(r(F)^.5))
46     if th==`j'
47 }
48 * Adjust testhour such that 9 starts at 9 (and is not centered at 9
49 gen testhour=th+7
50 replace testhour=9.375 if testhour==9
51 replace testhour=10.375 if testhour==10
52 replace testhour=11.375 if testhour==11
53 replace testhour=12.375 if testhour==12
54 replace testhour=13.375 if testhour==13
55 replace testhour=14.375 if testhour==14
56 * make tw plot
57 twoway (bar beta testhour, fcolor(orange_red) lwidth(medthick) lcolor(orange_red)
58 barwidth(.75)) ///
59         (rcap u l testhour, fcolor(black) lcolor(gs3) lwidth(medium)) ///
60         ,ylabel(#7 , nogrid noticks) graphregion(fcolor(white) lcolor(white)) ///
61         xlabel(,noticks) plotregion(fcolor(white) lcolor(black)) ///
62         ylabel(0,lcolor(black) ) title(" ") ///
63         ytitle("Change in test score (SD)") ///
64         legend(off) xtitle("Test starting time") ///
65         xlabel(8 "8:00" 9 "9:00" 10 "10:00" 11 "11:00" 12 "Noon" 13 "13:00")
66 graph export "$df\main_graph_testhour.png",replace width(4000)
67
68 /*****
69     2 The effect of breaks
70 *****/
71 use "$tf\analysisdata\sample'.dta",clear
72 * set fixed effect level
73 xtset instnr

```



```

73 * regress
74 set matsize 5000
75 preserve
76 * create dataset to save estimates
77 clear
78 set obs 7
79 gen id=_n
80 gen breakbeta=.
81 gen breakl=.
82 gen breaku=.
83 gen thbeta=.
84 gen thl=.
85 gen thu=.
86 save "$tf\estimates.dta",replace
87 restore
88 eststo clear
89 * program to run regressions
90
91 cap program drop myreg
92 program myreg
93     syntax, id(string) [condition(string)]
94     eststo: qui: xtreg testscore_std testhour break $controls2 `condition', cluster(
instnr) fe nonest
95     estadd scalar DF= e(df_m)
96     estadd scalar groups= e(N_g)
97     estadd scalar sgroup= e(g_min)
98     estadd scalar lgroup= e(g_max)
99     estadd scalar ar2=e(r2_a)
100     preserve
101     use "$tf\estimates.dta",clear
102     replace breakbeta=_b[break] if id==`id'
103     replace breakl=_b[break]-invttail(e(df_r),0.025)*_se[break] if id==`id'
104     replace breaku=_b[break]+invttail(e(df_r),0.025)*_se[break] if id==`id'
105     replace thbeta=_b[testhour] if id==`id'
106     replace thl=_b[testhour]-invttail(e(df_r),0.025)*_se[testhour] if id==`id'
107     replace thu=_b[testhour]+invttail(e(df_r),0.025)*_se[testhour] if id==`id'
108     save "$tf\estimates.dta",replace
109 end
110 * main
111 * singletons in instnr?
112 bys instnr: gen count=_n
113 bys instnr: egen instcount=max(count)
114 myreg, id(1) condition(if instcount>1)
115 myreg, id(2) condition(if sub==6 & instcount>1)
116 myreg, id(3) condition(if sub==1 & instcount>1)
117 myreg, id(4) condition(if grade<5 & instcount>1)
118 myreg, id(5) condition(if grade>5 & instcount>1)
119 myreg, id(6) condition(if breakdata==1 & instcount>1)
120 xtset id
121 drop count
122 bys id: gen count=_n
123 bys id: egen obs=max(count)
124 drop if obs==1
125 myreg, id(7)
126
127 * graph
128 use "$tf\estimates.dta",clear
129 gen expand=2
130 expand expand
131 bys id: gen show=_n
132 replace id=id-0.2 if show==1
133 replace id=id+0.2 if show==2
134 tw (bar breakbeta id if show==2, barwidth(0.4) fcolor(blue) lcolor(blue)) (rcap breaku
breakl id if show==2, lcolor(black) ) ///
135 (bar thbeta id if show==1, barwidth(0.4) fcolor(red) lcolor(red)) (rcap thu thl id
if show==1, lcolor(black)), ///
136 xlabel(1 "Main" 2 "Math" 3 "Reading" 4 "Young" 5 "Old" 6 "Breakdata" 7 "Ind. FE",
noticks) ///
137 yline(0,lcolor(black)) ylabel(-0.02(0.01)0.05, nogrid noticks) graphregion(fcolor(
white) lcolor(white)) ///
138 plotregion(fcolor(white) lcolor(black)) legend(order(3 "Hourly effect" 1 "Effect of
a break") region(lcolor(white))) xtitle("") ytitle("Effect size (SD)")
139 graph export "$df\main_breakdata.png",replace width(4000)
140 esttab using "$df\appendix_reg_hetero_table.csv", stats(groups sgroup lgroup DF ar2 aic N
, fmt(%11.3f)) b(%5.3f) ///

```

```

141         keep(break testhour) noline nonotes se nonumbers fragment ///
142         subs("[1em]" " ") label replace
143
144
145 /*****
146     3 figure break bars and quantile regs
147 *****/
148 * Load data
149
150 use "$tf\analysisdata\sample'.dta",clear
151 * set fixed effect level
152 xtset instnr
153 * singletons in instnr
154 bys instnr: gen count=_n
155 bys instnr: egen instcount=max(count)
156 drop if instcount==1
157 * regress
158 set matsize 5000
159 qui: xtreg testscore_std testhour break $controls2, cluster(instnr) fe
160 * create dataset to save estimates
161 preserve
162 clear
163 set obs 9
164 gen id=_n
165 gen u= .
166 gen l= .
167 gen betamean=_b[break]
168 gen betameanth=_b[testhour]
169 gen beta=.
170 gen uth=.
171 gen lth=.
172 gen betath=.
173 save "$tf\estimatesq.dta",replace
174 restore
175 * quantile reg
176 qui: xtreg testscore_std break $controls2,
177 predict fe, u
178 gen y=testscore_std-fe
179 qui:tab grade,gen(grade)
180 qui:tab sub,gen(sub)
181 qui:tab testyear,gen(testyear)
182 qui:tab dow,gen(dow)
183
184 forval i=1/9{
185     local l=`i'*10
186     qreg2 y break testhour gradel-grade6 subl-sub6 testyear1-testyear4 dowl-dow5 $covariates
187     $missings, quantile(`l') c(instnr)
188     preserve
189     use "$tf\estimatesq.dta",clear
190     replace u= _b[break]+invttail(e(df_r),0.025)*_se[break] if id==`i'
191     replace l= _b[break]-invttail(e(df_r),0.025)*_se[break] if id==`i'
192     replace beta=_b[break] if id==`i'
193     replace uth= _b[testhour]+invttail(e(df_r),0.025)*_se[testhour] if id==`i'
194     replace lth= _b[testhour]-invttail(e(df_r),0.025)*_se[testhour] if id==`i'
195     replace betath=_b[testhour] if id==`i'
196     save "$tf\estimatesq.dta",replace
197     restore
198 }
199
200 * MAKE GRAPH
201 use "$tf\estimatesq.dta",clear
202 replace id=id*10
203 tw (line beta id, lcolor(blue)) ///
204     (line betath id, lcolor(red)) ///
205     (line u id, lcolor(blue) lpattern(dash) ) ///
206     (line l id, lcolor(blue) lpattern(dash) ) ///
207     (line uth id, lcolor(red) lpattern(dash) ) ///
208     (line lth id, lcolor(red) lpattern(dash) ) ///
209     , xtitle(Testscore percentile) ytitle("Effect size (SD)") ///
210     legend(order ( 1 "Effect of a break" 2 "Hourly effect") ///
211     region(lcolor (white))) yline(0,lcolor(black)) ///
212     xlabel(0(10)100) ylabel(-0.03(0.01)0.05,nogrid noticks) graphregion(fcolor(white)
213     lcolor(white)) ///
214     xlabel(,noticks) plotregion(fcolor(white) lcolor(black))

```

```
214 graph export "$df\main_breakdata_quantiles.png",replace width(4000)
215
```

```

1  /*****
2  * Analyses for test time project, appendix
3  * Last edited: 20160113 by hhs@sfi.dk
4  *****/
5
6  * Load preamble
7  do "D:\Data\workdata\704335\Timeofday\dofiles\preamble.do"
8  * Set memory
9  set max_memory 4g, perm
10
11 /*****
12 * Outline:
13 * 1 Descriptives
14 * 2 Main regression table
15 * 3 Distributions
16 * 4 Compare precision
17 * 5 Main regression table with percentiles
18 * 6 Effect size
19 *****/
20
21
22
23
24 /*****
25 1 Descriptives
26 *****/
27 use "$tf\analysisdata\sample'.dta",clear
28
29 * My little program to create covariate means
30 cap program drop mytab
31 program mytab
32 syntax varlist [using/]
33 preserve
34     * close handle if open
35     cap file close mytab
36     * open handle
37     file open mytab using "`using'",write replace
38     * temporary variables
39     tempvar s1 s2 s3 s4 s5 s6 s7 s8
40     forval i=1/8{
41         qui: gen `s`i'=.
42     }
43     * Replace missing obs with missing, if relevant
44     cap replace `l'=. if missing_`l'==1
45
46     foreach l of local varlist{
47
48         * Save mean values
49         qui: sum `l'
50         qui: replace `s1'=r(mean)
51         qui: sum `l' if testhour==8
52         qui: replace `s2'=r(mean)
53         qui: sum `l' if testhour==9
54         qui: replace `s3'=r(mean)
55         qui: sum `l' if testhour==10
56         qui: replace `s4'=r(mean)
57         qui: sum `l' if testhour==11
58         qui: replace `s5'=r(mean)
59         qui: sum `l' if testhour==12
60         qui: replace `s6'=r(mean)
61         qui: sum `l' if testhour==13
62         qui: replace `s7'=r(mean)
63
64         local label: variable lab `l'
65         file write mytab "`label':" _tab _tab _tab %8.3f (`s1') ";" %8.3f (`s2') ";" %
8.3f (`s3') ";" ///
66                                     %8.3f (`s4') ";" %8.3f (`s5') ";" %8.3f (`s6') ";" %8.3f (`s7')
67     _n
68     }
69     * Number of Observations
70     forval i=1/1{
71         qui: sum testscore_std
72         qui: replace `s1'=r(N)
73         qui: sum testscore_std if testhour==8
74         qui: replace `s2'=r(N)

```

```

74     qui: sum testscore_std if testhour==9
75     qui: replace `s3'=r(N)
76     qui: sum testscore_std if testhour==10
77     qui: replace `s4'=r(N)
78     qui: sum testscore_std if testhour==11
79     qui: replace `s5'=r(N)
80     qui: sum testscore_std if testhour==12
81     qui: replace `s6'=r(N)
82     qui: sum testscore_std if testhour==13
83     qui: replace `s7'=r(N)
84 }
85     file write mytab "Observations:" _tab _tab _tab %12.0fc (`s1') ";" %12.0fc (`s2'
) ";" %12.0fc (`s3') ";" ///
86     %12.0fc (`s4') ";" %12.0fc (`s5') ";" %12.0fc (`s6') ";" %
12.0fc (`s7') ";" %12.0fc (`s8') _n
87     file close mytab
88     restore
89 end
90
91
92 * Main table of descriptives
93 mytab uncert schoolday $covariates $missings using "$df\appendix_descriptives.txt"
94 keep if breakdata==1
95 * Table of descriptives selected schools
96 mytab uncert schoolday $covariates $missings using
"$df\appendix_descriptives_surveyed.txt"
97
98 /*****
99     2 Main regression table
100 *****/
101 * Load data
102 use "$tf\analysisdata\sample'.dta",clear
103 * set fixed effect level
104 xtset instnr
105 * singletons in instnr
106 bys instnr: gen count=_n
107 bys instnr: egen instcount=max(count)
108 * Now create table with six edumms
109 eststo clear
110 * Test hour effect, no controls
111 eststo: qui: reg testscore_std testhour, cluster(instnr)
112 estadd scalar Fval=r(F)
113 estadd scalar Pval=r(p)
114 estadd scalar DF= e(df_m)
115 estadd scalar groups= e(N_g)
116 estadd scalar sgroup= e(g_min)
117 estadd scalar lgroup= e(g_max)
118 estadd scalar ar2=e(r2_a)
119 * Hourly effect, no controls
120 eststo: qui: reg testscore_std th2-th6, cluster(instnr)
121 qui: test th2=th3=th4=th5=th6=0
122 estadd scalar Fval=r(F)
123 estadd scalar Pval=r(p)
124 estadd scalar DF= e(df_m)
125 estadd scalar groups= e(N_g)
126 estadd scalar sgroup= e(g_min)
127 estadd scalar lgroup= e(g_max)
128 estadd scalar ar2=e(r2_a)
129 * Hourly effect, basic controls
130 eststo: qui: xtreg testscore_std th2-th6 $controls1 if instcount>1, cluster(instnr) fe
131 qui: test th2=th3=th4=th5=th6=0
132 estadd scalar Fval=r(F)
133 estadd scalar Pval=r(p)
134 estadd scalar DF= e(df_m)
135 estadd scalar groups= e(N_g)
136 estadd scalar sgroup= e(g_min)
137 estadd scalar lgroup= e(g_max)
138 estadd scalar ar2=e(r2_a)
139 * Hourly effect, extended controls
140 eststo: qui: xtreg testscore_std th2-th6 $controls2 if instcount>1, cluster(instnr) fe
141 qui: test th2=th3=th4=th5=th6=0
142 estadd scalar Fval=r(F)
143 estadd scalar Pval=r(p)
144 estadd scalar DF= e(df_m)
145 estadd scalar groups= e(N_g)

```



```

146 estadd scalar sgroup= e(g_min)
147 estadd scalar lgroup= e(g_max)
148 estadd scalar ar2=e(r2_a)
149 * Break effect, no controls
150 eststo: qui: reg testscore_std testhour break, cluster(instrnr)
151 estadd scalar DF= e(df_m)
152 estadd scalar groups= e(N_g)
153 estadd scalar sgroup= e(g_min)
154 estadd scalar lgroup= e(g_max)
155 estadd scalar ar2=e(r2_a)
156 * Break effect, basic controls
157 eststo: qui: xtreg testscore_std testhour break $controls1 if instcount>1, cluster(instrnr)
    fe
158 estadd scalar DF= e(df_m)
159 estadd scalar groups= e(N_g)
160 estadd scalar sgroup= e(g_min)
161 estadd scalar lgroup= e(g_max)
162 estadd scalar ar2=e(r2_a)
163 * Break effect, extended controls
164 eststo: qui: xtreg testscore_std testhour break $controls2 if instcount>1, cluster(instrnr)
    fe
165 estadd scalar DF= e(df_m)
166 estadd scalar groups= e(N_g)
167 estadd scalar sgroup= e(g_min)
168 estadd scalar lgroup= e(g_max)
169 estadd scalar ar2=e(r2_a)
170 esttab using "$df\appendix_reg_table.csv", stats( groups sgroup lgroup Fval Pval DF ar2
aic N, fmt(%11.3f)) b(%5.3f) ///
171     keep(break testhour th2 th3 th4 th5 th6) noline nonotes se nonumbers fragment ///
172     subs("[1em]" " ") label replace
173
174
175
176 /*****
177 3 Distribution
178 *****/
179 * Load data
180 use "$tf\analysisdata\sample'.dta",clear
181 * make density plots
182 twoway (kdensity testscore_std if testhour==8,lcolor(gs11) bwidth(0.25) kernel(triangle
)) ///
183 (kdensity testscore_std if testhour==9,lcolor(gs9) bwidth(0.25) kernel(triangle
)) ///
184 (kdensity testscore_std if testhour==10,lcolor(gs7) bwidth(0.25) kernel(triangle
)) ///
185 (kdensity testscore_std if testhour==11,lcolor(gs5) bwidth(0.25) kernel(triangle
)) ///
186 (kdensity testscore_std if testhour==12,lcolor(gs3) bwidth(0.25) kernel(triangle
)) ///
187 (kdensity testscore_std if testhour==13,lcolor(gs1) bwidth(0.25) kernel(triangle
)) ///
188 ,legend(order(1 "8AM" 2 "9AM" 3 "10AM" 4 "11AM" 5 "Noon" 6 "1PM") rows(2) ) ///
189 graphregion(fcolor(white) lcolor(white)) plotregion(fcolor(white) lcolor(black))
///
190 ylabel(,noticks) xlabel(,noticks) ytitle("Density") xtitle("Standardized test
score")
191 graph export "$df\appendix_test_distribution.png",replace width(4000)
192
193 use "$tf\analysisdata\sample'.dta",clear
194 * set fixed effect level
195 xtset instrnr
196 * singletons in instrnr
197 bys instrnr: gen count=_n
198 bys instrnr: egen instcount=max(count)
199 qui: xtreg testscore_std testhour break $controls2 if instcount>1, cluster(instrnr) fe
200
201 predict res, u
202 twoway (kdensity res if testhour==8,lcolor(gs11) bwidth(0.25) kernel(triangle)) ///
203 ,legend(off) ///
204 graphregion(fcolor(white) lcolor(white)) plotregion(fcolor(white) lcolor(black))
///
205 ylabel(,noticks) xlabel(,noticks) ytitle("Density") xtitle("Residuals")
206 graph export "$df\appendix_residual_distribution.png",replace width(4000)
207 /*****
208 4 Compare precision across samples

```

```

208 *****/
209 * Create empty dataset to save estimates
210 clear
211 set obs 4
212 gen id=_n
213 gen depvar="edu"
214 replace depvar="birthweight" if id==2
215 replace depvar="inc" if id==3
216 replace depvar="testscore_std" if id==4
217 gen expand=5
218 expand expand
219 drop expand
220 bys id: gen th=_n+1
221 gen expand=10
222 expand expand
223 drop expand
224 bys th depvar: gen sample=_n
225 gen samplesize=.
226 gen tstat=.
227 save "$tf\estimates.dta",replace
228 * Load data
229 use "$tf\analysisdata\sample'.dta",clear
230 xtset instnr
231 * singletons in instnr
232 bys instnr: gen count=_n
233 bys instnr: egen instcount=max(count)
234 drop if instcount==1
235 * run regressions
236 gen missing_testscore_std=testscore_std==.
237 foreach l in edu inc birthweight testscore_std {
238   forval a=1(1)10{
239     local i=`a'/10
240     di "`i'"
241     preserve
242     keep if runiform()<`i'
243     qui: sum testscore_std
244     local N=r(N)
245     qui: xtreg `l' th2-th6 $controls1 if missing_`l'!=1, cluster(instnr) fe
246     restore
247     preserve
248     use "$tf\estimates.dta",clear
249     replace samplesize=`N' if depvar=="`l'" & sample=="`a'"
250     forval j=2/6{
251       replace tstat=_b[th`j']/_se[th`j'] if depvar=="`l'" & sample=="`a'" & th=="`j'"
252     }
253     save "$tf\estimates.dta",replace
254     restore
255   }
256 }
257
258 use "$tf\estimates.dta",clear
259 * make graph
260 replace tstat=abs(tstat)
261
262
263 tw (scatter tstat samplesize if th==2 & depvar=="testscore_std",msymbol(o) mcolor(green
)) ///
264 (scatter tstat samplesize if th==3 & depvar=="testscore_std",msymbol(d) mcolor(green
)) ///
265 (scatter tstat samplesize if th==4 & depvar=="testscore_std", msymbol(s) mcolor(
green)) ///
266 (scatter tstat samplesize if th==5 & depvar=="testscore_std",msymbol(x) mcolor(green
)) ///
267 (scatter tstat samplesize if th==6 & depvar=="testscore_std",msymbol(t) mcolor(green
)) ///
268 (scatter tstat samplesize if th==2 & depvar=="edu",msymbol(o) mcolor(gs1)) ///
269 (scatter tstat samplesize if th==3 & depvar=="edu",msymbol(d) mcolor(gs1)) ///
270 (scatter tstat samplesize if th==4 & depvar=="edu", msymbol(s) mcolor(gs1)) ///
271 (scatter tstat samplesize if th==5 & depvar=="edu",msymbol(x) mcolor(gs1)) ///
272 (scatter tstat samplesize if th==6 & depvar=="edu",msymbol(t) mcolor(gs1)) ///
273 (scatter tstat samplesize if th==2 & depvar=="inc",msymbol(o) mcolor(gs7)) ///
274 (scatter tstat samplesize if th==3 & depvar=="inc",msymbol(d) mcolor(gs7)) ///
275 (scatter tstat samplesize if th==4 & depvar=="inc", msymbol(s) mcolor(gs7)) ///
276 (scatter tstat samplesize if th==5 & depvar=="inc",msymbol(x) mcolor(gs7)) ///
277 (scatter tstat samplesize if th==6 & depvar=="inc",msymbol(t) mcolor(gs7)) ///

```

```

278     (scatter tstat samplesize if th==2 & depvar=="b",msymbol(o) mcolor(gs12)) ///
279     (scatter tstat samplesize if th==3 & depvar=="b",msymbol(d) mcolor(gs12)) ///
280     (scatter tstat samplesize if th==4 & depvar=="b",msymbol(s) mcolor(gs12)) ///
281     (scatter tstat samplesize if th==5 & depvar=="b",msymbol(x) mcolor(gs12)) ///
282     (scatter tstat samplesize if th==6 & depvar=="b",msymbol(t) mcolor(gs12)) ///
283     (scatter tstat samplesize if th==22 & depvar=="b",msymbol(t) mcolor(white)) ///
284     ,legend(order(2 "Testscore" 7 "Parental Education" 12 "Parental income" 17 "Birth
weight" 21 " " 6 "9AM" 7 "10AM" 8 "11AM" 9 "Noon" 10 "1PM") rows(2) region(lcolor(white)))
    ///
285     ylabel(#7 ,nogrid noticks) graphregion(fcolor(white) lcolor(white)) ///
286     xlabel(,noticks) plotregion(fcolor(white) lcolor(black)) ///
287     ylabel("T-value (numerical)") xtitle(Sample size)
288
289
290
291     graph export "$df\appendix_graph_testhour_tstats_sample.png",replace width(4000)
292
293     /*****
294     5 Main regression table, with percentiles
295     *****/
296
297     * Load data
298     use "$tf\analysisdata\sample'.dta",clear
299     * set fixed effect level
300     xtset instnr
301     * singletons in instnr
302     bys instnr: gen count=_n
303     bys instnr: egen instcount=max(count)
304     eststo clear
305     * Test hour effect, no controls
306     eststo: qui: reg percentile testhour, cluster(instnr)
307     estadd scalar Fval=r(F)
308     estadd scalar Pval=r(p)
309     estadd scalar DF= e(df_m)
310     estadd scalar groups= e(N_g)
311     estadd scalar sgroup= e(g_min)
312     estadd scalar lgroup= e(g_max)
313     estadd scalar ar2=e(r2_a)
314     * Hourly effect, no controls
315     eststo: qui: reg percentile th2-th6, cluster(instnr)
316     qui: test th2=th3=th4=th5=th6=0
317     estadd scalar Fval=r(F)
318     estadd scalar Pval=r(p)
319     estadd scalar DF= e(df_m)
320     estadd scalar groups= e(N_g)
321     estadd scalar sgroup= e(g_min)
322     estadd scalar lgroup= e(g_max)
323     estadd scalar ar2=e(r2_a)
324     * Hourly effect, basic controls
325     eststo: qui: xtreg percentile th2-th6 $controls1 if instcount>1, cluster(instnr) fe
326     qui: test th2=th3=th4=th5=th6=0
327     estadd scalar Fval=r(F)
328     estadd scalar Pval=r(p)
329     estadd scalar DF= e(df_m)
330     estadd scalar groups= e(N_g)
331     estadd scalar sgroup= e(g_min)
332     estadd scalar lgroup= e(g_max)
333     estadd scalar ar2=e(r2_a)
334     * Hourly effect, extended controls
335     eststo: qui: xtreg percentile th2-th6 $controls2 if instcount>1, cluster(instnr) fe
336     qui: test th2=th3=th4=th5=th6=0
337     estadd scalar Fval=r(F)
338     estadd scalar Pval=r(p)
339     estadd scalar DF= e(df_m)
340     estadd scalar groups= e(N_g)
341     estadd scalar sgroup= e(g_min)
342     estadd scalar lgroup= e(g_max)
343     estadd scalar ar2=e(r2_a)
344     * Break effect, no controls
345     eststo: qui: reg percentile testhour break, cluster(instnr)
346     estadd scalar DF= e(df_m)
347     estadd scalar groups= e(N_g)
348     estadd scalar sgroup= e(g_min)
349     estadd scalar lgroup= e(g_max)
350     estadd scalar ar2=e(r2_a)

```

```

351 * Break effect, basic controls
352 eststo: qui: xtreg percentile testhour break $controls1 if instcount>1, cluster(instrnr) fe
353 estadd scalar DF= e(df_m)
354 estadd scalar groups= e(N_g)
355 estadd scalar sgroup= e(g_min)
356 estadd scalar lgroup= e(g_max)
357 estadd scalar ar2=e(r2_a)
358 * Break effect, extended controls
359 eststo: qui: xtreg percentile testhour break $controls2 if instcount>1, cluster(instrnr) fe
360 estadd scalar DF= e(df_m)
361 estadd scalar groups= e(N_g)
362 estadd scalar sgroup= e(g_min)
363 estadd scalar lgroup= e(g_max)
364 estadd scalar ar2=e(r2_a)
365 esttab using "$df\appendix_reg_table_percentiles.csv", stats( groups sgroup lgroup Fval
366 Pval DF ar2 aic N, fmt(%11.3f)) b(%5.3f) ///
367 keep(break testhour th2 th3 th4 th5 th6) noline nonotes se nonumbers fragment
368 ///
369 subs("[1em]" " ") label replace
370
371 /*****
372 6 Unconditional correlations for covariates and testscore
373 *****/
374 * Load data
375 use "$tf\analysisdata\sample.dta",clear
376
377 * regress
378 sum inc,d
379 replace inc=. if inc>r(p99)
380 replace inc=. if inc<r(p1)
381 sum birthweight,d
382 replace birthweight=. if birthweight>r(p99)
383 replace birthweight=. if birthweight<r(p1)
384 eststo clear
385 set matsize 5000
386 eststo: qui: reg testscore_std inc if missing_incrank!=1, cluster(instrnr)
387 estadd scalar ar2=e(r2_a)
388 eststo: qui: reg testscore_std birthweight if missing_bir!=1, cluster(instrnr)
389 estadd scalar ar2=e(r2_a)
390 eststo: qui: reg testscore_std edu if missing_edu!=1, cluster(instrnr)
391 estadd scalar ar2=e(r2_a)
392 eststo: qui: reg testscore_std schoold , cluster(instrnr)
393 estadd scalar ar2=e(r2_a)
394 esttab using "$df\appendix_main_effectsizes.csv", stats( ar2 aic N, fmt(%11.3f)) b(%6.4f
395 ) ///
396 keep( inc birthweight edu schoolday) noline nonotes se nonumbers fragment ///
397 subs("[1em]" " ") label replace

```